Avocados Vs. Millennials

Group 5

Adrian Lupercio
Jeremy Kalmus
Laura Gutierrez

Does the Millennial population affect the price of avocados?

Brief Summary

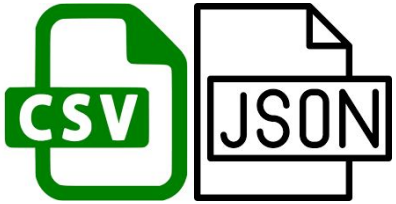
- Collected data
- Cleaned data
- Manipulated and merged data for better results
- Created charts and tables to represent this data
- Analyzed the result

Questions

- Does the millennial population percentage affect the price of avocados?
- Does the average household income affect the price of avocados?
- Does the weather affect the price of avocados?

Data

Types -



Sources -

OpenWeather

kaggle

United States® Census Bureau

Locations of Interest

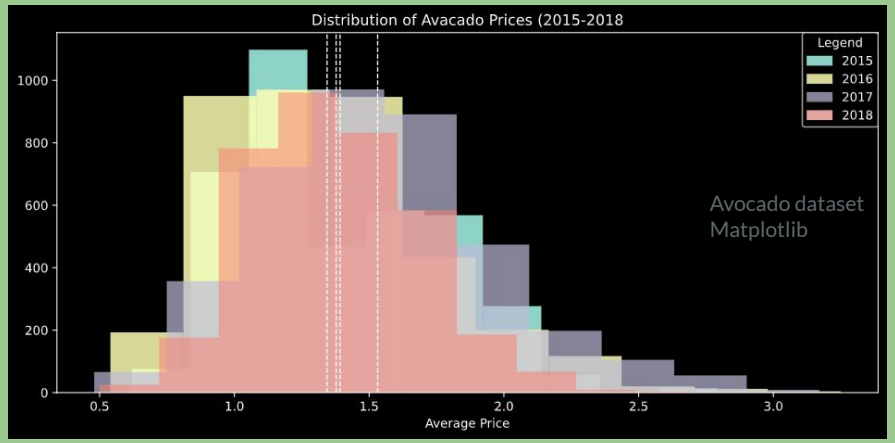
project_one > Laura > cities.csv

```
City, State
Albany, New York
Atlanta, Georgia
Boise City, Idaho
Boston, Massachusetts
Buffalo, New York
Charlotte, North Carolina
Chicago, Illinois
Cincinnati, Ohio
Columbus, Ohio
Dallas, Texas
Denver, Colorado
Detroit, Michigan
Grand Rapids, Michigan
Harrisburg, Philadelphia
Hartford, Conneticut
Houston, Texas
Indianapolis City, Indiana
Jacksonville, Florida
Las Vegas, Nevada
Los Angeles, California
Miami, Florida
Nashville, North Carolina
New Orleans, Louisiana
New York, New York
Orlando, Florida
Philadelphia, Philadelphia
Phoenix, Arizona
Pittsburgh, Philadelphia
Portland, Oregon
Raleigh, North Carolina
Richmond, Virgina
Roanoke, Virgina
Sacramento, California
San Diego, California
San Francisco, California
Seattle, Washington
Spokane, Washington
St. Louis, Missouri
Syracuse, New York
Tampa, Florida
```

Avocado data

13 columns, 18722 Rows
Average Price

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
2	1/4/15	1.22	40873.28	2819.5	28287.42	49.9	9716.46	9186.93	529.53	0	conventional	2015	Albany
3	1/11/15	1.24	41195.08	1002.85	31640.34	127.12	8424.77	8036.04	388.73	0	conventional	2015	Albany
4	1/18/15	1.17	44511.28	914.14	31540.32	135.77	11921.05	11651.09	269.96	0	conventional	2015	Albany
5	1/25/15	1.06	45147.5	941.38	33196.16	164.14	10845.82	10103.35	742.47	0	conventional	2015	Albany
6	2/1/15	0.99	70873.6	1353.9	60017.2	179.32	9323.18	9170.82	152.36	0	conventional	2015	Albany
7	2/8/15	0.99	51253.97	1357.37	39111.81	163.25	10621.54	10113.1	508.44	0	conventional	2015	Albany
8	2/15/15	1.06	41567.62	986.66	30045.51	222.42	10313.03	9979.87	333.16	0	conventional	2015	Albany
9	2/22/15	1.07	45675.05	1088.38	35056.13	151	9379.54	9000.16	379.38	0	conventional	2015	Albany
10	3/1/15	0.99	55595.74	629.46	45633.34	181.49	9151.45	8986.06	165.39	0	conventional	2015	Albany
11	3/8/15	1.07	40507.36	795.68	30370.64	159.05	9181.99	8827.55	354.44	0	conventional	2015	Albany
12	3/15/15	1.11	43045.79	2128.26	30447.17	99.67	10370.69	9989.59	381.1	0	conventional	2015	Albany
13	3/22/15	1.12	46346.85	2141.83	34313.56	141.8	9749.66	9252.6	497.06	0	conventional	2015	Albany
14	3/29/15	1.02	67799.08	1402.28	58623.22	89.5	7684.08	7208.49	475.59	0	conventional	2015	Albany



Cleaning Data

- Removed excess and unrelated information.
- Removed years not in the 3 main data sets.
- Formatted the context ie city names to be the same across data sets.

Normal Test results for 2015 are statistic=3.165418348345291 with a pvalue=0.2054178309964272)

Normal Test results for 2016 are statistic=4.917488597178161 with a pvalue=0.08554229914031346)

Normal Test results for 2017 are statistic=8.92214851212411, with a pvalue=0.011549949017861348)

Normal Test results for 2018 are statistic=1.8896928899913943, with a pvalue=0.3887392591775346)

Data USA

API and JSON to pull data for the 42 cities in Avocado Dataset

2 CSV Files

1) Median Income
   a) 3 Columns, 1476 Rows
2) Population By Age Range
   a) 4 columns, 16236 Rows

Further Cleaning:

-Drop Duplicates
-Join Table for City and City, State

```
for geo in geography:
    print(geo)

    query_1_url = 'https://datausa.io/api/data?measures=Birthplace,Birthplace%20Moe&drilldowns=Age&{}'.format(geo)
    results_1 = requests.get(query_1_url).json()

    for year in years:
        for i in range(len(results_1['data'])):
            try:
                results_1 = requests.get(query_1_url).json()
                city_data_1["City"].append(results_1['data'][i]['Geography'])
                city_data_1["Year"].append(results_1['data'][i]['Year'])
                city_data_1['Age Range'].append(results_1['data'][i]['Age'])
                city_data_1["Number in Range"].append(results_1['data'][i]['Birthplace'])
            except:
                pass
```

```
#query 2 loop for getting income data
years = [2013,2014,2015,2016,2017,2018]

city_data_2 = {
    "Year": [],
    "City": [],
    "Household Income by Race": []
}
counter = 0
for geo in geography:
    counter += 1
    print(counter)
    query_2_url = 'https://datausa.io/api/data?measure=Household%20Income%20By%20Race,Household%20Income%20By%20Race%20Moe&{}:parents,16000US0644000,16000US0644000:similar'.format(geo)



    results_2 = requests.get(query_2_url).json()

    for year in years:
        for i in range(len(results_2['data'])):
            try:
                results_2 = requests.get(query_2_url).json()
                city_data_2['City'].append(results_2['data'][i]['Geography'])
                city_data_2['Household Income by Race'].append(results_2['data'][i]['Household Income by Race'])
                city_data_2['Year'].append(results_2['data'][i]['Year'])
            except:
                pass
```

	A	B	C	D
1		Year	City	Median Income
2	0	2018	Albany, NY	45500
3	1	2017	Albany, NY	43790
4	2	2016	Albany, NY	42335
5	3	2015	Albany, NY	40949
6	4	2014	Albany, NY	41099
7	5	2013	Albany, NY	40287

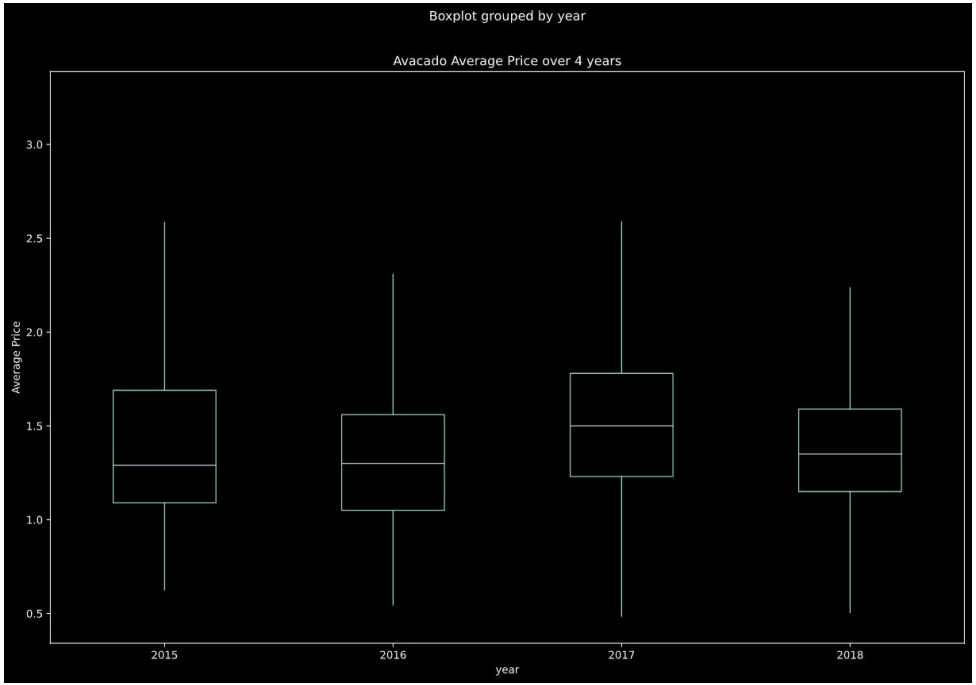
Median Income

	A	B	C	D	E
1		Year	City	Age Range	Number in Range
2	0	2018	Albany, NY	18 to 24 Years	21091
3	1	2018	Albany, NY	75 Years & Over	5906
4	2	2018	Albany, NY	45 to 54 Years	10514
5	3	2018	Albany, NY	Under 5 Years	5491
6	4	2018	Albany, NY	65 to 74 Years	6670

Age Group

Analysis 1: Income Vs. Average Price

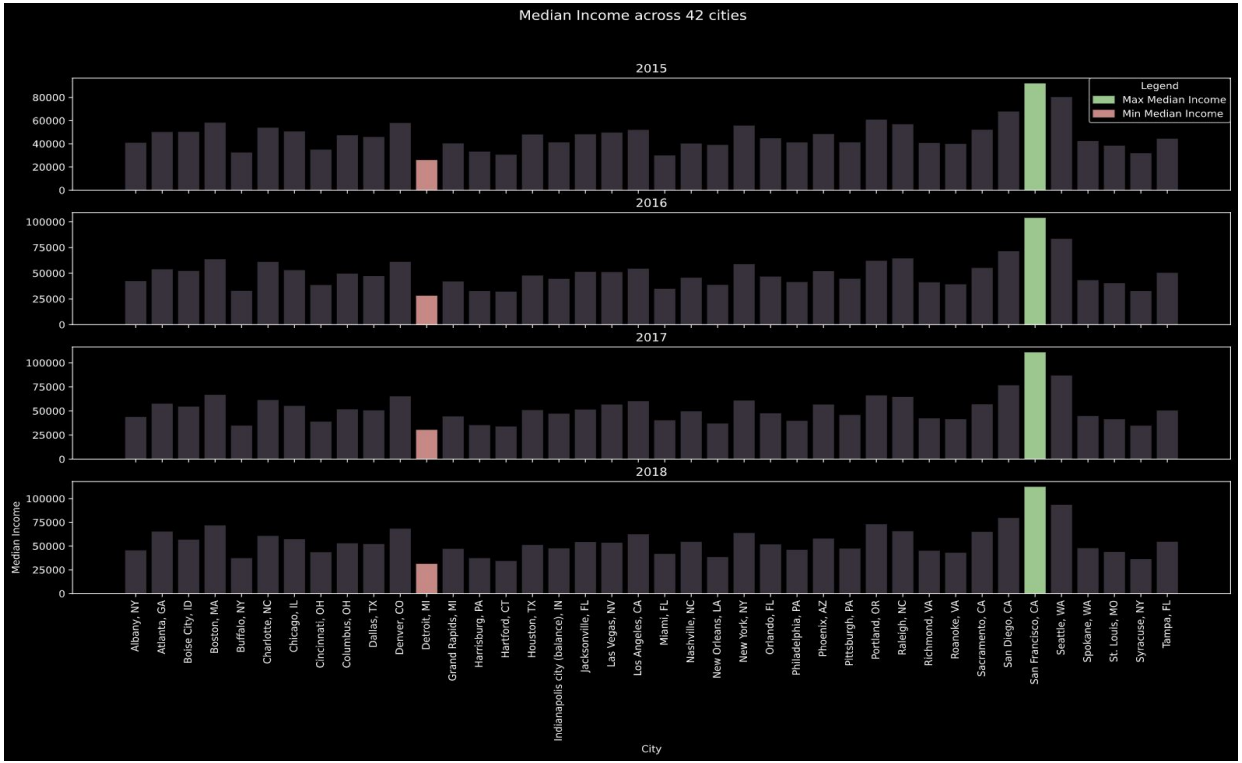
Analysis: Average Price



	AveragePrice
count	15170.000000
mean	1.411760
std	0.384503
min	0.480000
25%	1.120000
50%	1.370000
75%	1.660000
max	3.250000

Avocado dataset
Pandas

Analysis 1: Income Vs. Average Price



Median Income with Highs and Lows Highlighted
plt.subplots()

Income dataset
Matplotlib

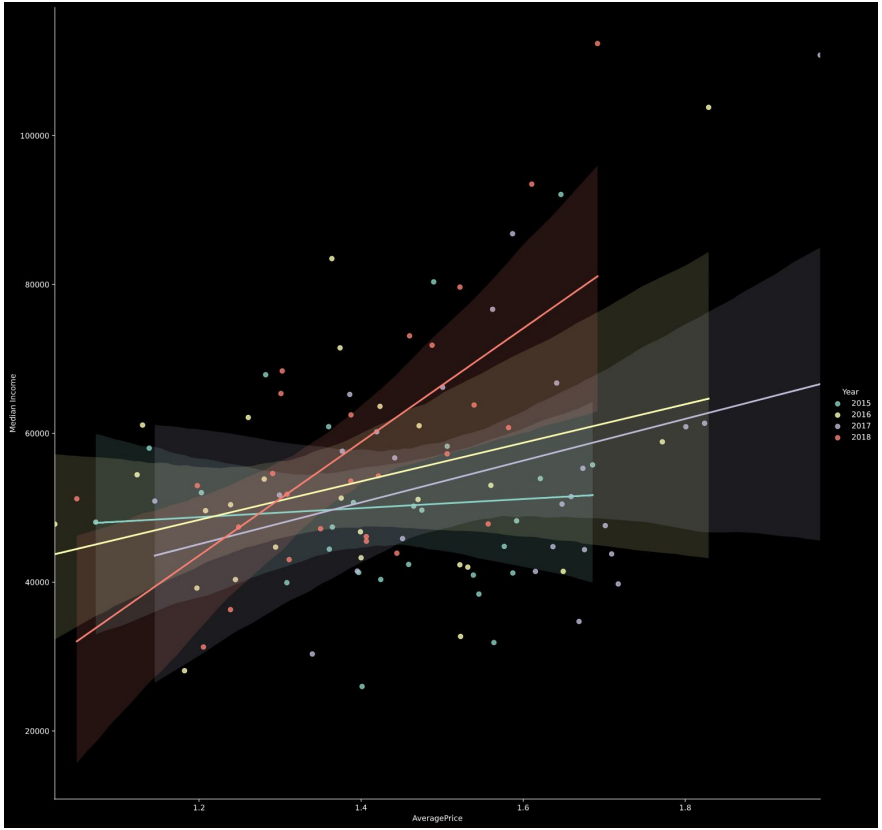
Analysis 1: Income Vs. Average Price

The line equation for 2015 is y = 6076.81x + 41442.05
The r-Squared for 2015 is 0.004458147154015406

The line equation for 2016 is y = 25900.41x + 17284.68
The r-Squared for 2016 is 0.10472518139494967

The line equation for 2017 is y = 28068.11x + 11416.46
The r-Squared for 2017 is 0.09341235056496099

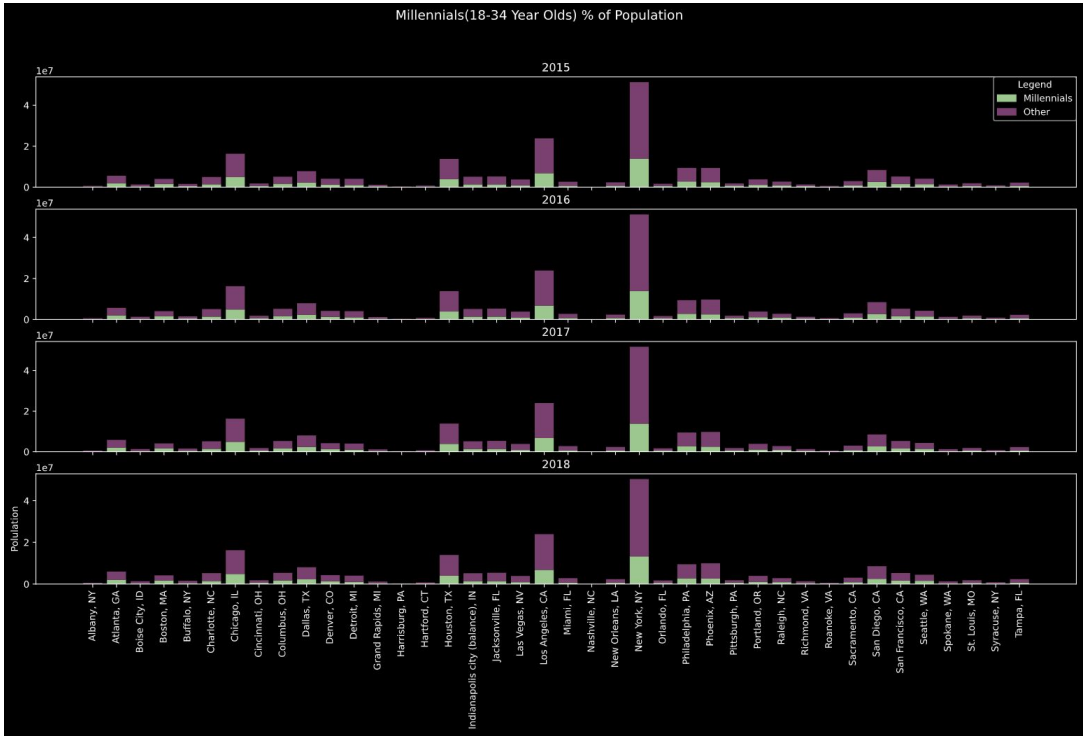
The line equation for 2018 is y = 76348.44x + -48057.92
The r-Squared for 2018 is 0.4123616254754726



Avocado dataset & Median Income dataset
Seaborn

Analysis 2: Millennial Populations Vs. Average Price

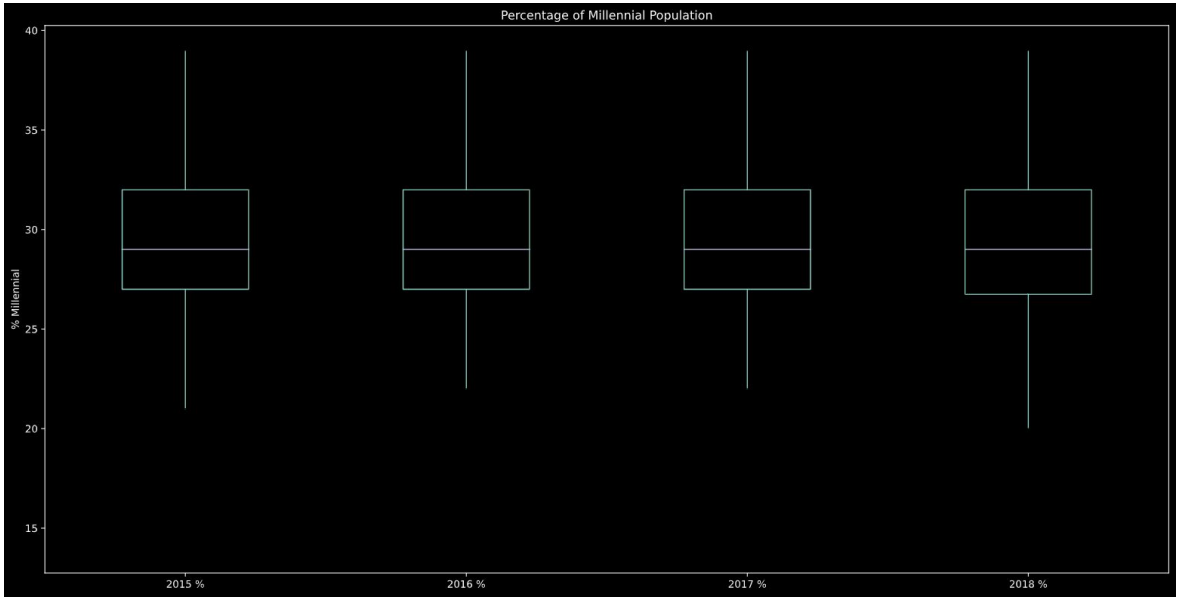
Analysis 2: Millennial Populations Vs. Average Price



Age group dataset
np.where[]
Matplotlib

City	2015 %	2016 %	2017 %	2018 %
Albany, NY	38.55%	38.89%	38.75%	37.84%
Atlanta, GA	34.15%	34.32%	34.09%	34.65%
Boise City, ID	26.38%	26.17%	26.17%	26.07%
Boston, MA	38.93%	39.34%	39.28%	39.44%
Buffalo, NY	29.61%	30.60%	30.95%	29.62%
Charlotte, NC	27.37%	27.23%	27.90%	27.56%
Chicago, IL	30.01%	29.86%	30.01%	29.80%
Cincinnati, OH	31.81%	32.03%	32.25%	32.07%
Columbus, OH	32.31%	32.03%	32.09%	32.37%
Dallas, TX	28.18%	28.86%	29.28%	28.17%
Denver, CO	30.88%	30.96%	30.99%	31.33%
Detroit, MI	24.92%	25.46%	25.54%	25.53%
Grand Rapids, MI	32.11%	32.66%	32.98%	33.49%
Harrisburg, PA	27.59%	27.87%	28.32%	28.78%
Hartford, CT	32.11%	31.96%	31.81%	31.30%
Houston, TX	28.55%	28.02%	27.98%	28.32%
Indianapolis city (balance), IN	26.26%	26.54%	26.54%	26.58%
Jacksonville, FL	25.67%	25.80%	25.74%	25.51%
Las Vegas, NV	22.50%	22.32%	22.29%	22.34%
Los Angeles, CA	28.56%	28.23%	28.12%	28.40%
Miami, FL	24.55%	24.23%	24.95%	23.67%
Nashville, NC	20.70%	14.06%	15.38%	19.99%
New Orleans, LA	28.29%	27.94%	27.12%	26.38%
New York, NY	27.18%	27.01%	26.78%	26.32%
Orlando, FL	33.38%	32.56%	31.35%	32.22%
Philadelphia, PA	29.37%	29.33%	29.15%	28.95%
Phoenix, AZ	25.64%	25.48%	25.58%	26.46%
Pittsburgh, PA	38.34%	36.94%	38.55%	37.57%
Portland, OR	28.27%	27.99%	27.92%	27.19%
Raleigh, NC	31.31%	31.76%	31.71%	31.02%
Richmond, VA	34.64%	34.49%	34.38%	34.25%
Roanoke, VA	23.89%	23.57%	23.45%	23.16%
Sacramento, CA	27.63%	28.02%	26.46%	27.81%
San Diego, CA	30.58%	30.86%	30.91%	29.30%
San Francisco, CA	30.64%	30.62%	30.61%	30.55%
Seattle, WA	34.43%	34.02%	33.00%	34.55%
Spokane, WA	27.50%	26.95%	26.99%	26.60%
St. Louis, MO	29.84%	29.73%	29.43%	29.19%
Syracuse, NY	34.55%	34.27%	34.03%	33.99%
Tampa, FL	27.69%	26.74%	26.98%	28.22%

Analysis 2: Millennial Populations Vs. Average Price



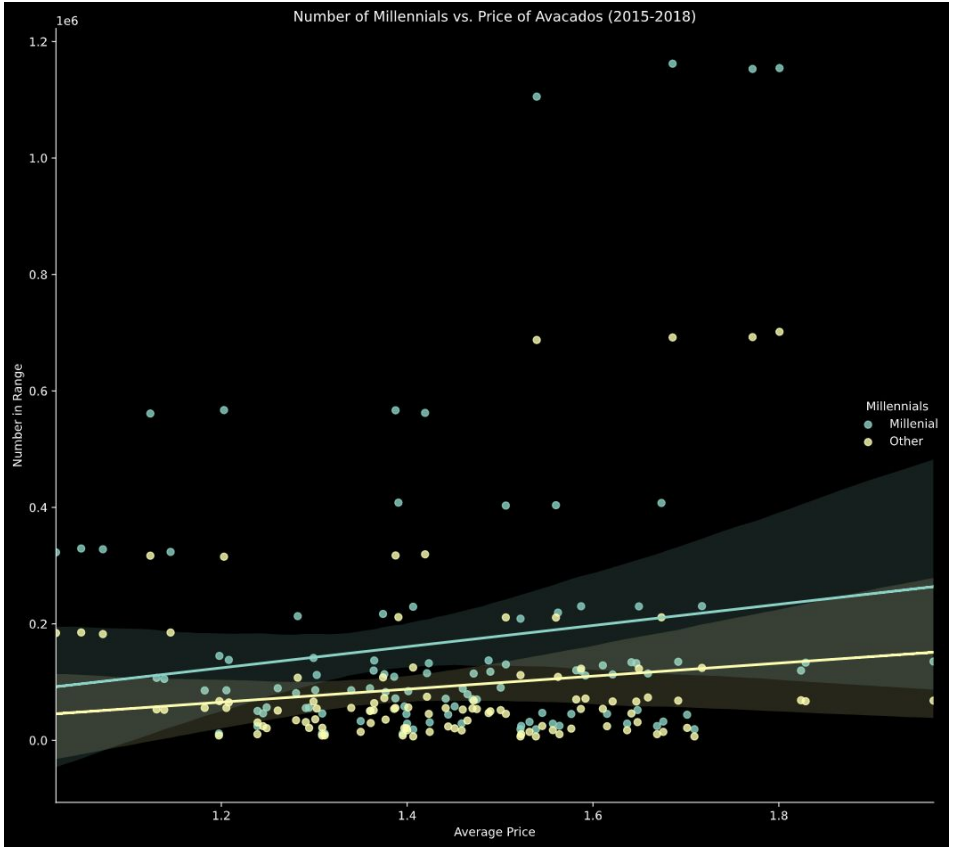
	2015 %	2016 %	2017 %	2018 %
count	40.000000	40.000000	40.000000	40.000000
mean	29.700000	29.400000	29.375000	29.400000
std	4.121240	4.650889	4.600376	4.241432
min	21.000000	14.000000	15.000000	20.000000
25%	27.000000	27.000000	27.000000	26.750000
50%	29.000000	29.000000	29.000000	29.000000
75%	32.000000	32.000000	32.000000	32.000000
max	39.000000	39.000000	39.000000	39.000000

Age Group dataset
Pandas

Analysis 2: Millennial Populations Vs. Average Price

The line equation for Millenial is y = 181922.13x + -93868.77
The r-Squared for Millenial is 0.021035153654571443

The line equation for Other is y = 111806.21x + -68659.15
The r-Squared for Other is 0.02217732375560323



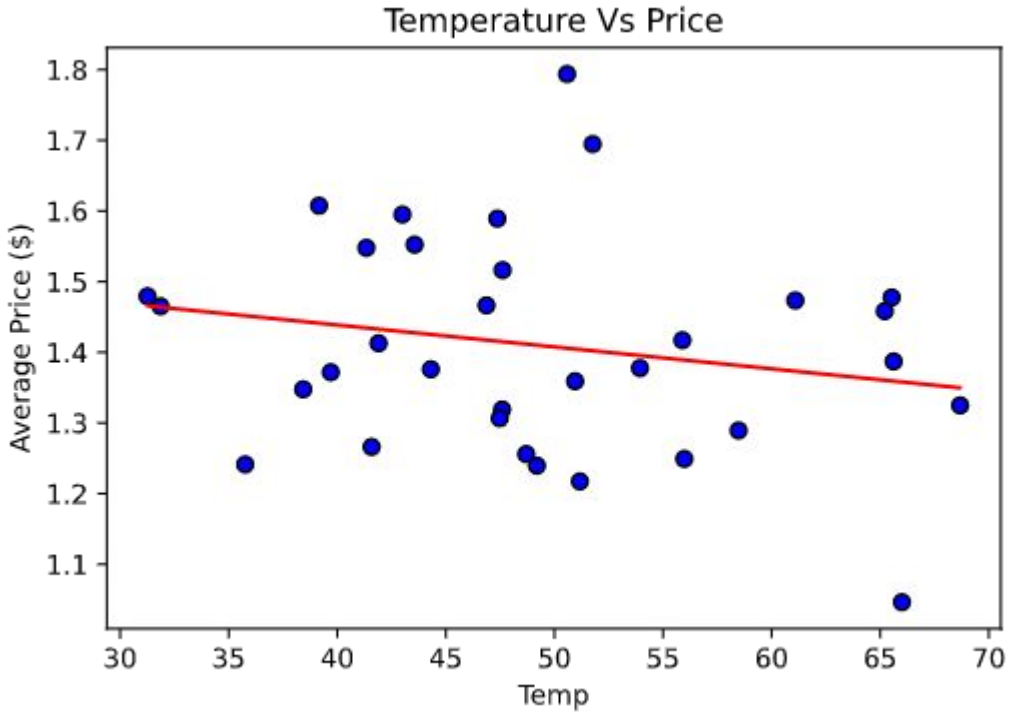
Avocado dataset & Age Group dataset
Seaborn

Analysis 3: Temperature Vs. Average Price

Does the weather in cities affect the price?

The line equation is y = -0.0x + 1.56

The r-Squared is 0.040818288930768776



Findings

Does the millennial population percentage affect the price of avocados?

Does the average household income affect the price of avocados?

Does the weather affect the price of avocados?

Other Considerations

- Distance from Harvest to purchase location
- Harvest yield on any given year
- Time of year and availability of Avocados
- Many other factors

Questions?